DOCUMENT RESUME

ED 453 268 TM 032 805

AUTHOR Camburn, Eric; Correnti, Richard; Taylor, James
TITLE Examining Differences in Teachers' and Researchers' Understanding of an Instructional Log.
SPONS AGENCY Department of Education, Washington, DC.; National Science Foundation, Arlington, VA.
PUB DATE 2001-04-00
NOTE 53p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Evaluation Methods; Factor Analysis; *Journal Writing; Language Arts; Reading Comprehension; *Reading Instruction; Self Evaluation (Individuals); *Teachers; *Validity

ABSTRACT

This study, part of the "Study of Instructional Improvement," assessed the validity of quantitative measures of reading comprehension instruction derived from a language arts instructional log. The log was designed to measure the language arts instructional experiences individual students received over time, and measures from the log were to be used as independent variables in statistical models of student achievement growth. Data are from two nested studies. In the first study, 91 teachers completed language arts instructional logs every school day for a 3-month period. In a subsidiary study, a subset of 31 of these teachers were observed by 2 researchers on 1 day of instruction during the 3 months. This study yielded 3 independent ratings (teacher and 2 observers) of 23 reading comprehension lessons. Measures of reading comprehension instruction were constructed from the first study's data set using factor analysis, and then proxy measures of reading comprehension instruction were created from the observations from the second study. Initial assessments of the validity of the reading comprehension measure were made using two indicators that gauged inter-rater agreement in the second study. The validity of the measures was further studied by evaluating qualitative evidence form the observations. The factor analyses of the data from the first study provide fairly traditional quantitative evidence of the construct validity of the measures. The measures of inter-rater agreement go beyond to give a general indication of the degree to which there was shared understanding of what occurred during the classroom observations. The three analyses yielded markedly different pictures of the validity of the reading comprehension instruction measures, including the fact that teachers generally reported much higher levels of reading comprehension activity than observers, and that raters marked different items about one quarter of the time. Appendixes contain the Language Arts Log and a glossary for the log. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 453 268

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
E. Camburn
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Examining Differences in Teachers' and Researchers' Understanding of an Instructional Log

Eric Camburn, Richard Correnti, and James Taylor

The University of Michigan

This paper was presented at the annual meeting of the American Educational Research Association, Seattle, WA, April, 2001. The research reported herein was conducted as part of the *Study of Instructional Improvement* which is supported by the U.S. Department of Education, the National Science Foundation, and several private funders. The authors wish to acknowledge the efforts of Sally Atkins-Burnett at the University of Michigan for her advice on the measures of reading comprehension instruction. Please address correspondence to: Eric Camburn, University of Michigan, School of Education, 610 East University, Room 3112A, Ann Arbor, MI 48109.

TM032805

BEST COPY AVAILABLE

Overview

This paper assesses the validity of quantitative measures of reading comprehension instruction derived from a language arts instructional log (a copy of the log can be found in Appendix A). The log is designed to measure the language arts instructional experiences individual students receive over time. Measures from the log will be used as independent variables in statistical models of student achievement growth.

We used multiple sources of evidence and multiple methods to seek out convergent and divergent evidence in order to evaluate the trustworthiness of the reading comprehension instruction measures. Our data come from a set of two nested studies. In the first study, the *Instructional Log Pretest* (referred to hereafter as the *Pretest*), 91 teachers completed Language Arts instructional logs every school day for a three month period . The *Instructional Log Validation Study* (referred to hereafter as the *Validation Study*) was a subsidiary study in which a subset of 31 of the 91 teachers were observed by two researchers on one day of instruction during the three month period. The *Validation Study* yielded three independent ratings (teacher and two observers) of 23 reading comprehension lessons. For a more detailed description of the *Validation Study* data and methodology, please see Barnes, Correnti, Taylor and Atkins-Burnett (2001).

Measures of reading comprehension instruction were constructed from the *Pretest* dataset using factor analysis. Using the same items used for the *Pretest* measures, proxy measures of reading comprehension instruction were created from the *Validation Study* observation dataset. Initial assessments of the validity of the reading comprehension measures were made using two indicators that gauge inter-rater agreement in the *Validation Study*. The validity of the measures was further investigated by evaluating three sources of qualitative evidence from the *Validation Study*: narrative reports of observed lessons, post-observation interviews with teachers, and

observers' notes in which they reflected upon discrepancies between their reports of a lesson and those of their co-observer. A number of themes emerged during our investigation of the qualitative evidence that help explain inter-rater differences.

The analyses presented in this paper are part of an ongoing effort to gauge how well we are measuring the instructional experiences of students and to more generally improve the measurement of instruction. Taken as a whole, these analyses provide rich diagnostic information that will help us improve measures of reading comprehension instruction derived from the logs. But we believe the analyses also provide insight into differential pictures of validity one obtains when different methods are used. The factor analyses of the *Pretest* data provide fairly traditional quantitative evidence of the construct validity of the measures. The measures of inter-rater agreement go a step beyond the fairly superficial factor analysis results to give a general indication of the degree to which there is shared understanding of what occurred during the classroom observations that were part of the *Validation Study*. The qualitative analyses go yet another layer deeper and help illuminate why teachers' and observers' impressions of classroom instruction during the observations differed. These three analyses yielded markedly different pictures of the validity of the reading comprehension instruction measures.

Orienting Perspectives

Our approach for this paper was to investigate validity from multiple perspectives using a variety of methods. This orientation led us to choose two primary modes of inquiry to organize and orient our work: 1) we primarily focused on inter-rater agreement as a measure of validity,

and 2) we sought both convergent and divergent evidence of agreement.[1] Each mode of inquiry, along with key ideas that oriented our work are discussed below.

We Sought Both Convergent and Divergent Evidence

Our overall goal was to better understand the validity of our measures rather than prove our measures were valid. To that end we sought both convergent and divergent evidence. Convergent evidence provides more convincing support for a measure or inference than a single rating of an entity and divergent evidence often leads one to revise hypotheses or to an enhanced understanding of a construct being studied (Jick, 1979). Therefore, by either confirming pre-conceptions or by identifying problems in pre-conceptions, seeking and evaluating both kinds of evidence can ultimately give researchers greater confidence in the conclusions they reach.

Our primary evidence of validity is inter-rater agreement

This paper primarily focused on inter-rater agreement as an indicator of measure validity. We interpreted inter-rater agreement as an indicator of the degree of shared understanding between teachers and researchers about analytic constructs being measured. Inter-rater agreement is often treated as an issue of reliability. For example, LeCompte and Goetz (1982) describe internal reliability as "the extent to which sets of meanings held by multiple observers are sufficiently congruent so that they describe phenomena in the same way" (p.41). However, Moss (1994) contends that evidence about inter-rater reliability can inform construct validity arguments:

> "Although the focus here is on reliability (consistency among independent measures intended as interchangeable), it should be clear that reliability is an aspect of construct

[1] Assessments of the validity of the instructional logs are not limited to the analyses presented in this paper. In the larger study we have sought out several contributing lines of evidence to establish the validity of these instruments. In an attempt to insure that the logs tap important subject matter a number of content standards documents, student assessments, and other questionnaires were reviewed. The instruments were also reviewed by a number of reading and mathematics experts. In addition, through a series of pilot studies, we solicited input from teachers about how well their log reports represented what they did in their classrooms. Taking these steps during the instrument development process helped us assess the content, construct, and face validity of the logs.

validity (consonance among multiple lines of evidence supporting the intended interpretation over alternative interpretations)."

In other words, if independent audiences have high inter-rater agreement, then their scores and, perhaps, interpretations of events and how to code those events may be interchangeable. Such evidence puts us on the path toward convergent evidence supporting the validity of a measure since multiple independent observers arrive at a similar interpretation. Convergent evidence is attained upon confirmation that independent raters reached similar interpretations through a similar *process*.

Cognitive psychologists who study how individuals respond to survey items also lend support to the contention that the degree of shared understanding between researchers and research subjects is an important determinant of validity. Such scholars generally agree that the first step in answering an item is to understand its meaning (Sudman, Bradburn and Schwarz, 1995) . Research has shown that "lexical ambiguity", which occurs when words potentially take on more than one meaning for a respondent, is a common problem in question comprehension. Lexical ambiguities are inherent in everyday language. One reason for this is that alternative meanings of the same word may be differentially accessible to different people because of the frequency with which they use or encounter the word. Lexical ambiguity can also result when words take on different meanings within different groups and subcultures. In the context of this study, teachers and researchers potentially represent groups that might have different understandings of instructional terms used on the logs (Freeman, 1996). Ambiguous meaning is a clear threat to valid measurement since the validity of a response is dependent on respondents sharing researchers' understanding of an item's meaning. Because of their heavy reliance upon technical terms, instructional log items may be particularly prone to problems associated with lexical ambiguity.

In using inter-rater agreement as an indicator of measure validity, we were guided by the following perspectives and assumptions: 1) we did not privilege one rater's reports over another; 2) however, we did expect that raters' perspectives on an observation and their subsequent log reports would significantly depend on whether they were a teacher or a researcher; and 3) we expected that disagreement among raters would partially be a function of unique features of the logs. Each of these perspectives are discussed separately below.

We did not privilege one rater's reports over another. For each classroom observation conducted as part of the *Validation Study* , we obtained three independent ratings of the instruction provided to students --one rating from the teacher being observed, and one rating apiece from each of two observers. If we had assumed observers' ratings as a standard, comparing teachers' and observers' ratings would have been more akin to a test of criterion validity, where the observers' ratings were used as the criterion. Because asymmetries existed in the knowledge that observers and teachers brought to the observation and in the roles they played during the observation, we were uncomfortable holding up the reports of a particular kind of rater as a criterion. This orientation led us to create quantitative indices of inter-rater agreement that summarized the reports of all three raters. Reflecting this orientation, we refer to teachers and observers generically as "raters."

We expected teachers and observers to perceive observed instruction differently, and for those differences in perception to be reflected in log reports. One important area of difference we expected to see was in the background knowledge raters brought with them to the observation. We expected that knowledge of the following things would significantly color raters' reports: of the particular students in the class, of students in general, of teaching, of the curriculum being used, of the teacher's intentions during the lesson, of how the teacher typically

teaches reading comprehension, and of the researchers' definitions of terms used in the log. As should be clear from this list, we believe an asymmetry exists in raters' background knowledge in areas that would directly affect how a rater would interpret an episode of classroom instruction, and subsequently, how that interpretation would be reflected in a log report. Obviously teachers' knowledge of their students, of the curriculum they're using, and generally of their teaching of reading comprehension will be immensely superior to that of observers. *Validation Study* observers on the other hand may have had a better understanding of terms used in the log because they underwent a much more rigorous training on the instrument than teachers (Barnes, Correnti, and Taylor, 2001). We also anticipated that inter-rater agreement would partly be a function of the different roles raters performed during observations. During observations, teachers devoted their attention to teaching , not to tracking or categorizing what they were doing. In contrast, observers' only responsibility during instructional events was to observe and record what teachers and students were doing. Because of this, teachers may have missed some events that happened with the target student because their attention was elsewhere.

We expected that inter-rater agreement would partially be a function of unique features of the logs. Compared to standard surveys the logs are extraordinarily complex. The logs contain a large number of items and the items capture very discrete facets of instruction, very often using technical terminology. Given the large number of discrete aspects of instruction being measured, the logs contain redundant items that make it possible to check different, but similar items. We believe that in some instances, these redundancies may have amplified between-rater differences. Regarding the language used on the log, lexical ambiguity has been shown to adversely impact the accuracy of survey reports (Lackner and Garrett, 1972; Fee, 1979) and we worried that some of the language on the log might be misunderstood by teachers. We

attempted to ameliorate this problem by providing teachers and observers with a glossary of terms used in the logs (see Appendix B for the log glossary).

These orienting perspectives drove our choice of analytic strategies, and our analysis and interpretation of the evidence available to us. Because we applied multiple methods to a wide variety of evidence sources, we chose to integrate descriptions of our methods within discussions of the results produced by each method.

Results

Three sets of analyses were conducted for this paper. We first constructed measures of reading comprehension instruction whose validity we wished to evaluate. In the second set of analyses, we assessed the reading comprehension instruction measures using indices of inter-rater agreement. In the final set of analyses, we examined qualitative evidence to search for explanations of observed levels of disagreement between raters.

Creating Measures of Reading Comprehension Instruction

Measures of reading comprehension instruction were created in three steps: 1) we conducted exploratory factor analysis of *Pretest* log data to identify constructs and item clusters that could be used to measure the constructs, 2) we evaluated the construct validity of the item clusters by analyzing their factor structure and other statistical evidence, and 3) we used the item clusters to create proxy measures from *Validation Study* instructional log data. Each of these steps is discussed further below.

One of the most unique features of the instructional log is that it is completed multiple times throughout a school year for an individual student. Consequently, measures derived from the logs are based on a series of reports across a year. In contrast, log data from the *Validation Study* were gathered from only a single day of instruction. Ideally, the design of the *Validation*

Study would have yielded multiple observations per student just like the larger study. This was not possible however, given cost constraints and the need to limit time demands on teachers. Acknowledging this limitation, we looked for other ways to incorporate the aspect of multiple observations over time into the reading comprehension instruction measures. We were able to do this by using the larger *Pretest* dataset to define the constructs and choose the items used for these measures.

Investigators of the Study of Instructional Improvement had identified, a priori, a number of instructional constructs they felt were being measured by the instructional log. In order to confirm these constructs and identify others being measured by the log, the reading comprehension instruction items were subjected to exploratory factor analysis. Items used to measure reading comprehension instruction come from two sections of the instructional log which contain a combined total of 38 items. The items in these sections are measured as dichotomies in which teachers indicate whether they did or did not do something on a particular day. Since factor analysis requires variables to be measured at the interval or ratio level (Kim and Mueller, 1978), we decided to aggregate students' daily log reports to weekly reports. This had the effect of putting the log items onto an interval scale while still maintaining multiple observations per student and a reasonable unit of analysis.

The initial extraction from a principal components analysis of the reading comprehension instruction items sections yielded 12 factors with eigenvalues greater than one. These factors accounted for 62.7 percent of the variance in the items. The factors were rotated using a Varimax rotation and the factors in the resulting component matrix were evaluated for their theoretical sensibility. Items with factor loadings greater than .45 were considered part of the item cluster comprising a factor. Of the initial 12 factors, a subset of six were identified as

being conceptually sensible. In an effort to attempt to bring greater conceptual and empirical clarity to the factors, a second principal components analysis was run, this time forcing a six factor solution. This solution accounted for 44.1 percent of the variance in the 38 reading comprehension instruction items. Again the rotated component matrix of factor loadings was evaluated. For this paper we decided to evaluate the validity of measures based on three of the six rotated factors that were the most theoretically sensible. As a final step, separate factor analyses which specified a single factor solution, were run for each of the three item clusters. Again items were assigned to a cluster if they had a factor loading of .45 or greater. Tables 1-3 display the results of these separate analyses and Cronbach's alpha for the three sets of items.

In tables 1-3 and throughout this paper, we make reference to log item numbers which begin with the letter "B." These numbers are printed on the log itself thus allowing readers to reference the full text of items analyzed in this paper (see Appendix A).

Table 1: Cluster 1 - Summarize or describe text that was read

Item	Factor loading
Worked on summarizing (B1d)	.738
Identifying main ideas and details (B1n)	.594
Provide extended oral answers (B2d)	.535
Practice teacher-selected comprehension strategy (B2j)	.589
Conduct a thinkaloud (B2l)	.568
Frame and ask questions about text (B2m)	.504
Retell story (B2o)	.584
Summarize text (B2p)	.722
Percent of variance explained by this factor	37.149
Cronbach's Alpha	.751

Items in cluster 1 appear to be measuring post-reading activity in which students are asked to summarize or characterize what they have read and then relate that to the teacher. Many of the items indicate that this often happens through verbal exchange between teachers and

students. It is our interpretation that when these items are checked, students are generally required to be more expansive and to go into greater depth than they are in more truncated question/answer exchanges with teachers or with short answer worksheet exercises. "Identifying main ideas and details" is thought to be a subsidiary activity in that students are commonly asked to summarize or describe main ideas or details of the text, and that identifying such information is often a precursor to summarization and description.

Table 2: Cluster 2 - Guided comprehension strategies

Item	Factor loading
Activating prior knowledge (B1a)	.725
Clarifying, monitoring for meaning (B1b)	.643
Making personal connections to story or text (B1d)	.647
Making predictions, generating questions (B1e)	.752
Previewing, surveying (B1f)	.451
Listen to text read to them (B2a)	.744
Provide brief oral comments or answers (B2b)	.759
Justify answers or explain reasoning (B2n)	.637
Percent of variance explained by this factor	45.783
Cronbach's Alpha	.829

The items in cluster 2 are thought to capture a number of explicit pre- and during-reading strategies that students are using to make sense of text they are reading. We refer to these as "guided" strategies, because they often take place within the context of teacher-directed activity such as the teacher reading a story, or the teacher asking questions.

Table 3: Cluster 3 - Comprehension review skills

Item	Factor loading
Locating answers or information	.702
Comparing, contrasting	.644
Drawing conclusions, making inferences	.729
Provide brief written comments or answers	.683
Provide extended written comments or answers	.562
Complete sentences with correct word or words	.602
Percent of variance explained by this factor	43.07
Cronbach's Alpha	.731

The items in cluster 3 appear to measure a set of post-reading activities in which students convey discrete details about a text, often in a written format. A number of the items tap the ways in which students were asked to retrieve and process this information prior to conveying it -- i.e. they first locate answers or information, and once located, students compare or contrast story elements or they draw conclusions about the information they located. The information itself, and students' analyses of it is ultimately conveyed to the teacher through brief written comments or answers, or through some sort of written exercise.

The factor analysis results suggest that in most cases, individual items that make up a cluster are substantially correlated with an underlying dimension that ties the items together. Loadings for all factors except one were greater than .5. The Alpha reliability statistics reported for these factors suggests that teachers' cumulative weekly log reports are quite internally consistent. That is, within a particular week, teachers frequently checked items within a cluster together. Because the reports used in the factor analyses were weekly aggregates, temporal simultaneity in the original log reports was lost, thus, we are very cautious in our interpretation of Cronbach's Alpha in this case.

By a number of traditional standards, these empirical results suggest that the item clusters may be tapping constructs that have some salience in the classroom. The correlational patterns among the items seem theoretically sensible and teachers seem to have answered the items in a fairly consistent manner. However, the percent of variance explained by these clusters is relatively low, indicating that it is likely that these items are also measuring factors other than those covered in this paper.

Based on these factor analyses, the three item clusters presented above were used to create proxy measures of reading comprehension instruction from the *Validation Study* data. We refer to the measures as proxies because they differ from the measures that will ultimately be created for the larger study in two important ways. First, measures in the larger study will be based on multiple reports per student, whereas measures from the *Validation Study* are based on reports from a single day of instruction. Second, measures in the larger study will be constructed using an item response theory (IRT) model which is a more rigorous measurement model than was used here. The proxy measures used for this paper were created by simply taking the mean of the items within a cluster. Since the log items are dichotomous, a rater's "score" on a measure is simply the proportion of items within a cluster he/she checked. We limited our analyses to 23 of the 31 *Validation Study* observations in which a teacher report of reading comprehension instruction was available.

Indices of Inter-rater Agreement

We sought to produce an index that reflected our orientation of not privileging one rater's reports over another. This led us to create indices that holistically characterized the degree of inter-rater agreement for a particular observation. Operationally, we assessed the degree of inter-rater agreement by measuring its opposite, the degree to which raters disagreed in their ratings.

The first index, which we refer to as the *measure disagreement index(MDI)*, captures the degree to which raters of an observation differed on the proxy measures. The index was created by simply summing the absolute value of the differences in raters' proxy measure scores for every possible rater pair. Values on the index are thus the average difference between raters' proxy measures for a given observation. The formula for this index is given below:

Measure disagreement index

$$\frac{\sum_{p=1}^{P_{ij}} \left| \overline{X_i} - \overline{X_j} \right|}{P_{ij}}$$

for p=1 to P rater pairs, i=teacher, teacher, observer 1; and j=oberserver 1, observer 2, observer 2

Since the proxy measures are comprised of multiple items, we were cognizant of the possibility that raters could have equal scores on a measure while choosing different items within a cluster. This possibility suggested that in general, between-rater differences in measures would not necessarily be synonymous with between-rater differences in item selections. To probe this possibility we created a second index called the *item disagreement index (IDI)*. We expected that distinguishing between score-level and item-level disagreement would not only provide us with a richer picture of between-rater differences generally, but would provide a check on the validity of the *measure disagreement index*. Specifically, we felt that this was an important check because in cases where item-level disagreement was greater than measure-level disagreement, the *measure disagreement index* would under-represent the degree of inter-rater disagreement. The *item disagreement index* was created by summing the absolute value of the differences in raters' item choices for every possible pair of rater responses to K items within an item cluster:

Item disagreement index

$$\frac{\sum_{p=1}^{P_{ijk}} \left| X_{ik} - X_{jk} \right|}{P_{ijk}}$$

for p=1 to P rater pairs; i=teacher, teacher, observer 1; j=oberserver 1, observer 2, observer 2; and k=1 to K items.

Since log items are dichotomous, the result of any paired comparison between raters on a single item can be either zero or one (i.e. either they chose the same item or they did not). Thus, values on the *item disagreement index* indicate the proportion of item selections associated with an observation in which raters disagreed. Tables 4-6 display proxy measure, MDI, and IDI scores for each *Validation Study* observation. Observations in these tables are sorted by their value on the *measure disagreement index*.

To gain a better understanding of how the indices work, it is instructive to look at a couple of individual observations. Consider observations for Mrs. Armstrong, Ms. Jacoby and Mr. Derr from table 4. In all three cases all raters had the exact same score on the "summarize/describe what was read" measure. The measure disagreement index reflects this with a value of zero for these observations. In the case of Ms. Jacoby, all raters agreed that none of the activities falling into this measure occurred. The item disagreement index score of zero for Mr. Derr indicates that not only did all three raters have the same scores on this measure, to produce those scores, the three raters chose exactly the same items. In contrast, despite the fact that their scores were equal, some of the items chosen by the raters for Mrs. Armstrong were different, leading to an item disagreement index score of .170. Clearly, simply examining inter-rater agreement on measure scores does not provide a complete picture of validity.

Consider another example, Ms. Getty(2) from the same table. Observer 2 and the teacher had the same scores on this measure. Observer 1's score on the other hand was substantially lower. These two facts are reflected in the MDI score. Specifically, despite the fact that two

raters had matching scores on the measure, one rater was so far from the other two that it produced a fairly high MDI score. The IDI is also quite high for this observation, indicating that on average, raters chose different items about half the time. These results strongly suggest that it would be useful to probe the observer narratives for Ms. Getty's instruction further to try to understand why the raters saw things so differently.

With a better understanding of how the indices work in hand, we now look more generally at what the indices tell us. One of the first things one notices is that the averages on the MDI is similar for all three measures. The average difference between the scores of any two raters on the proxy measures is about .18. The level of item disagreement varied by measure. For the guided comprehension strategy measure, raters chose different items an average of 32% of the time. For the other two measures, raters chose different items about one quarter of the time. It was striking that there were only three instances overall in which all three raters chose the exact same items. It was equally striking that the mean score for teachers on the three measures is nearly always higher than the means for the two observers. In other words, teachers consistently reported more reading comprehension activities than observers. These differences are particularly striking for the "summarize/describe what was read" measure.

The overall levels of inter-rater disagreement on the proxy measures strikes us as moderately high. The observations with the highest level of disagreement seem potentially problematic to us and clearly merit further scrutiny as to why raters in these observations differed so substantially. This analysis strongly suggests that the factor analysis results yield an overly positive portrayal of the validity of the reading comprehension instruction measures.

Table 4: Rater Scores and Disagreement Indices for "Summarize/describe what was read" Measure

	Summarizing/describing what was read				
Observation[2]	*Teacher*	*Observer 1*	*Observer 2*	*Measure disagreement index*	*Item disagreement index*
Ms. Becker[3]	0.250	0.250		0.000	0.500
Mrs. Armstrong(1)[4]	0.130	0.130	0.130	0.000	0.170
Ms. Jacoby	0.000	0.000	0.000	0.000	0.000
Mr. Derr	0.250	0.250	0.250	0.000	0.000
Mrs. Bartolo	0.130	0.250	0.250	0.080	0.250
Mrs. Armstrong(2)	0.130	0.130	0.000	0.080	0.170
Ms. Antos	0.000	0.130	0.130	0.080	0.080
Ms. Temple	0.130	0.000	0.000	0.080	0.080
Ms. Kritchfield	0.250	0.130	0.000	0.170	0.250
Mrs. Carter(1)	0.500	0.250	0.250	0.170	0.170
Mrs. Carter(2)	0.500	0.250	0.250	0.170	0.170
Ms. Petri	0.250	0.500	0.250	0.170	0.170
Mrs. Roberts	0.250	0.250	0.000	0.170	0.250
Mr. Rothchild	0.000	0.130	0.250	0.170	0.250
Ms. Page	0.000	0.380	0.250	0.250	0.250
Ms. Getty(1)	0.500	0.130	0.130	0.250	0.250
Ms. Getty(2)	0.500	0.130	0.500	0.250	0.500
Ms. Karsten	0.500	0.250	0.130	0.250	0.330
Ms. Stevenson	0.380	0.000	0.000	0.250	0.250
Ms. Gentry	0.630	0.130	0.250	0.330	0.330
Ms. Carroll	0.750	0.380	0.250	0.330	0.330
Ms. Booth	0.750	0.250	0.250	0.330	0.330
Ms. Sawyer	0.630	0.000		0.630	0.630
Mean	0.322	0.187	0.168	0.183	0.248
s.d	0.244	0.130	0.133	0.145	0.152

[2] The names listed on tables 4-6 and throughout this paper are pseudonyms.
[3] There was only one observer in this case
[4] Three teachers--Armstrong, Carter, and Getty--reported on two separate students apiece. The dataset contains two unique observations for these teachers which are denoted with (1) and (2) respectively.

Table 5: Rater Scores and Disagreement Indices for Guided Comprehension Strategies Measure

	Guided comprehension strategies				
Observation	*Teacher*	*Observer 1*	*Observer 2*	*Measure disagreement index*	*Item disagreement index*
Ms. Becker	0.500	0.500	.	0.000	0.250
Mrs. Bartolo	0.500	0.500	0.630	0.080	0.170
Ms. Page	0.380	0.500	0.380	0.080	0.080
Mrs. Armstrong(2)	0.750	0.630	0.630	0.080	0.330
Ms. Kritchfield	0.380	0.380	0.250	0.080	0.330
Mrs. Armstrong(1)	0.750	0.630	0.630	0.080	0.170
Ms. Petri	0.250	0.250	0.130	0.080	0.250
Ms. Karsten	0.630	0.500	0.500	0.080	0.420
Ms. Antos	0.380	0.500	0.500	0.080	0.330
Ms. Jacoby	0.500	0.500	0.630	0.080	0.330
Ms. Gentry	0.380	0.250	0.500	0.170	0.170
Ms. Carroll	0.880	0.630	0.630	0.170	0.250
Ms. Getty(1)	0.630	0.380	0.380	0.170	0.500
Ms. Getty(2)	0.630	0.380	0.500	0.170	0.500
Mr. Rothchild	0.500	0.250	0.500	0.170	0.250
Ms. Temple	0.250	0.500	0.500	0.170	0.250
Mrs. Carter(2)	0.750	0.380	0.500	0.250	0.250
Mrs. Roberts	0.630	0.500	0.250	0.250	0.330
Ms. Stevenson	0.880	0.750	0.500	0.250	0.420
Mrs. Carter(1)	0.750	0.380	0.250	0.330	0.330
Ms. Booth	0.000	0.380	0.500	0.330	0.330
Mr. Derr	0.750	0.130	0.380	0.420	0.420
Ms. Sawyer	0.750	0.130	.	0.630	0.630
Mean	0.557	0.432	0.460	0.183	0.317
s.d	0.223	0.159	0.145	0.142	0.125

Table 6: Rater Scores and Disagreement Indices for Comprehension Review Skills Measure

	Comprehension review skills				
Observation	*Teacher*	*Observer 1*	*Observer 2*	*Measure disagreement index*	*Item disagreement index*
Mrs. Roberts	0.330	0.330	0.330	0.000	0.000
Ms. Gentry	0.170	0.000	0.170	0.110	0.110
Mrs. Bartolo	0.000	0.170	0.170	0.110	0.110
Mrs. Armstrong(2)	0.330	0.330	0.500	0.110	0.220
Mrs. Armstrong(1)	0.330	0.330	0.500	0.110	0.220
Ms. Petri	0.170	0.170	0.330	0.110	0.330
Ms. Karsten	0.500	0.330	0.500	0.110	0.440
Ms. Antos	0.170	0.170	0.000	0.110	0.110
Mr. Derr	0.170	0.000	0.000	0.110	0.110
Ms. Stevenson	0.330	0.170	0.170	0.110	0.110
Ms. Becker	0.170	0.000	.	0.170	0.170
Ms. Sawyer	0.170	0.330	.	0.170	0.500
Ms. Page	0.000	0.330	0.170	0.220	0.220
Ms. Getty(1)	0.330	0.000	0.000	0.220	0.220
Ms. Getty(2)	0.330	0.000	0.000	0.220	0.220
Mrs. Carter(1)	0.670	0.500	0.330	0.220	0.220
Mrs. Carter(2)	0.830	0.500	0.500	0.220	0.330
Ms. Booth	0.500	0.330	0.170	0.220	0.220
Ms. Jacoby	0.670	0.500	0.330	0.220	0.220
Mr. Rothchild	0.500	0.670	0.330	0.220	0.440
Ms. Carroll	0.670	0.170	0.330	0.330	0.330
Ms. Kritchfield	0.330	0.330	0.830	0.330	0.440
Ms. Temple	0.170	0.500	0.670	0.330	0.440
Mean	0.341	0.268	0.301	0.177	0.249
s.d	0.221	0.192	0.227	0.084	0.135

Qualitative Evidence Bearing on Inter-rater Agreement

In order to better understand the validity of the reading comprehension measures more generally, and the statistical results in particular, we examined qualitative evidence from the *Validation Study*. Documents from three sources of qualitative evidence --observers' narrative observation reports, teacher follow-up interview transcripts, and reflective notes on between-researcher discrepancies--were placed into a single NUD*IST database. All source documents were coded in terms of specific references to individual log items. For example, if a teacher asked students to relate the main idea of a story to her and this was reflected in observers'

narrative notes and the teacher referred to this portion of the lesson in her follow-up interview, the appropriate sections of the classroom narratives and teacher interview were coded "B1n", the item number for "identifying main ideas and details." The reading comprehension items used for this paper comprise only about one quarter of all log items. Because of this, we decided to limit our analyses to the portions of the qualitative source documents that referenced the subset of items on which our measures are based. The extraction of text that was referenced to items in our three measures was achieved through a Boolean node search in NUD*IST.

Our first step in qualitative analysis was to read this extracted text, looking primarily for explanations of inter-rater disagreement. As we read, preliminary themes were noted. Next, using the MRI and IRI scores from tables 4-6, we identified observations we thought would be particularly informative. We re-read the extracted text for this subset of observations, going back to original source documents for greater contextual information or clarification when necessary. As we read through this subset of observations a second time we refined the list of preliminary explanations of inter-rater disagreement. Based on this second reading, summaries of individual observations were produced. These summaries included a brief description of the key instructional events in the observation and a brief summarization of the readers' hunches about what accounted for the degree of inter-rater disagreement present in the observation. Finally, we evaluated these summaries and synthesized the main explanations for inter-rater disagreement that seemed to be supported by the qualitative evidence.

In evaluating qualitative evidence from the *Validation Study* observations, five reasons for inter-rater disagreement emerged: 1) oversights on the part of raters, 2) teachers and observers interpreted terms used in log items differently, 3) observers lacked crucial contextual information that teachers possessed, 4) raters differed in their interpretation of the significance of

an instructional event, and 5) item redundancies in the log led to inconsistent responses from raters. As may be apparent from a cursory review of this list, the reasons for inter-rater disagreement were often complex and not easily compartmentalized. Consequently, there were a number of cases for which more than one reason from this list were applicable. When this occurred, we attempted to take note. Our discussion of the qualitative findings presented below is organized around the five reasons for inter-rater disagreement.

Oversights. Raters' log reports often disagreed because one or more raters failed to mark a log item because of an oversight. In an observational study, subjective decisions about what to attend to and what to observe will naturally introduce error into measures produced from the observations. Within the context of the *Validation Study*, teachers seemed particularly susceptible to this sort of error because they were simultaneously teaching and observing.

We found that simple oversights were a common cause of discrepancy, especially when the instruction occurred in a more peripheral context such as a test, a game or computer instruction. As one observer put it, "The other observer recognized a cloze procedure on the computer language arts instruction. I either didn't see it, or saw it but didn't code it." The observer notes that she may not have seen the instruction in question and her narrative indeed does not reflect it. We refer to this as an *observation oversight*. However, the observer also notes the possibility of seeing something, recording it in their observation notes, but then not coding this piece of instruction on the log. This scenario represents a second kind of oversight which we refer to as a *coding oversight*. For example, the same observer states, "The other observer caught that Mrs. K. contrasted the story they were reading to Cinderella. I missed it – remembered it happening, [but] didn't connect it to a code." In fact, despite the following entry

in this observer's notes, the instruction corresponding to this segment was not recorded on the log.

> Mrs. Kritchfield: Near the beginning of the story, Sonia called Annette her wicked stepmom. What other fairy tale also had a wicked stepmom?
>
> S: Sleeping beauty?
>
> Mrs. Kritchfield responded that this was not the right answer, then elicited the right answer (Cinderella) from another student."

This sort of coding oversight was usually associated with segments of instruction that were especially brief, often as brief as a single short question (see discussion of raters' perception of the significance of an instructional event below for further elaboration). Another example of a coding oversight occurs in the case of Ms. Karsten. For this observation, both observers clearly documented in their narratives that students were asked to "Complete sentences with the correct word or words" (item B2r). However, neither observer marked this item, despite the fact that they both included the following questions from the worksheet in their narrative:

1) Theodore could not meet his cousin at the forest because_____

2) Theodore could not walk so_____

Although they were generally less prone to oversights than observers, teachers were far from immune to the problem. Ms. Booth's curious omission of items B2a and B2b is an example of a teacher's coding oversight. From the notes of both observers it is clear that students in Ms. Booth's class read a text in a small group and that oral exchanges around the text occurred. However, while both observers marked items B2a (Student listened to text read to them) and B2b (Student provided brief comments or answers - oral), the teacher did not. It's not clear whether Ms. Booth's answers were due to an oversight, or whether she truly felt her log answers reflected what she did in class. If the former is true, then her responses may be

considered random measurement error, an inherent part of any survey. If the latter is true, then Ms. Booth's answers may signal a more significant problem.

Curiously, a total of six teachers omitted the fact that students were read to during their comprehension work. These teachers failed to mark B2a "listen to text read to them", while both observers marked that item. The following is taken from the observation of one of those teachers, Ms. Getty:

> At 11:04, Ms. Getty begins a taped version of the story on an old, difficult to hear tape recorder. The tape begins with some very odd computerized music, then a soothing woman's voice begins to read, starting by introducing the book's title, author and illustrator. The tape is a straight read-through of the story as it appears in the book. Children are to follow along in their own books as they listen to the tape.

On possible reason for the omission is that Ms. Getty's students were supposed to be reading along with the tape, and she may have thought B2s "students read assigned text" was a more appropriate item. However, that item was not marked either. Ms. Getty indicated that students were using a literature based or thematic short selection (B3e), that she read to the target student...asking minimal comments or questions (B4c) and that she listened to the target student read to her...asking minimal comments or questions (B4f). Given the relative richness of Ms. Getty's log report, her omission of B2a appears to be a coding omission.

Teachers' and observers' interpreted log terminology differently. Perhaps the single most prevalent reason for inter-rater disagreement were discrepancies in how raters interpreted terminology used in log items. Despite our attempts to promote a shared understanding of log terminology by providing a glossary and thorough training, raters still utilized very subjective definitions of many terms. The most common theme we saw was that a number of teachers tended to adopt more general, common-sense, or intuitive interpretations of terms while many observers applied much more specific interpretations of log terminology that tended to be more in line with glossary definitions. We believe these differences in interpretation are in part related

to differences in the knowledge and experiences of raters. For example, observers had more extensive training on the log and glossary, greater interaction with log developers, and it is likely that they had greater motivation to use the glossary in a meticulous manner. Likewise, knowledge of the content and instructional activities contained on the log likely varied considerably from teacher to teacher, and this in turn, was thought to color teachers' responses and their use of the glossary.

The items measuring whether students summarized or described a text after they read it were particularly susceptible to multiple interpretation by raters. In most cases, the teacher marked several more items within this measure than did either observer. This is reflected in the fact that the mean score for teachers on this measure (.32) was considerably higher than those of observer 1 and observer 2 (.19 and .17, respectively). Looking at two cases helps illustrate the kinds of situations that led to these differences.

When Ms. Carroll was observed, all three raters agreed on two items (B1n and B2j) in the "summarize/describe what was read" measure. Ms. Carroll also marked B1g "Summarizing", B2l "Thinkaloud", B2m "Students framed and asked questions" and B2o "retell story." One of the observers did not mark any of these items while the other observer marked only B1g from this set. The main difference between raters seems to revolve around Ms. Carroll's conception of student discourse and its nuances.

In her follow-up interview Ms. Carroll offered the following descriptions of "conduct a thinkaloud" and "retell story" (B2l and B2o). In order to get a sense of contrast, the corresponding glossary definitions for these items are listed after the excerpts from Ms. Carroll's interview.

Conduct a thinkaloud

> From Ms. Caroll's follow-up interview: "For the thinkaloud, that was discussing, like, after they read something, what they thought it meant. Why they thought it meant that, that's where they get into making connections, I think and talking about their connections."
>
> Glossary definition: The target student orally discussed or explained their thinking about and attempts to understand a story or text as they read it.

Retell story

> From Ms. Caroll's follow-up interview: Interviewer: "For 'retell story,' what were you thinking about? Ms. Caroll: When we were talking, we were just basically going over and when they were doing their worksheet on the problem and the solution that was really retelling what happened in the story."
>
> Glossary definition: The target student generated an account of a story that includes both important (major) and minor details. If the target student provides a succinct account of the main features of a story, record this activity in "Summarize text."

One basic distinction between Ms. Caroll's account of a thinkaloud and the glossary definition is the timing within the reading process when the thinkaloud occurred. While Ms. Carroll's students were "thinking aloud" after they had read a text, the glossary definition indicates that thinkalouds are a "during reading" activity. The glossary definition further specifies that thinkalouds are metacognitive exercises in which students verbalize how they are attempting to make sense of a text. Even though Ms. Caroll said her students were asked to demonstrate "why they thought it meant that", observation notes from the lesson segment to which she was referring indicated that students were being asked to share discrete details about the text through brief question-and-answer exchanges with Ms. Caroll. Thus the observation notes and Ms. Caroll's interview suggest that her students were verbally recounting details of a text they had read rather than verbalizing how they were coming to understand a text while they read it. Ms. Caroll's interpretation of a thinkaloud was not altogether unique. Ms. Booth described her students' use of thinkalouds as follows: "They were thinking out loud. They were

talking to each other and telling them what they thought about the story or what the story was about."

In the case of Mrs. Jaeger, one observer differed from the other two raters because he interpreted the item B1k "comparing, contrasting" much more narrowly. Mrs. Jaeger's follow-up interview and observation notes from both observers all reference a lesson segment in which students were asked to compare and contrast three houses. According to the notes of one observer, Mrs. Jaeger even used the word comparison in class as she interacted with a student:

> "So would you write a comparison between your house and grandmother's house and how that is different from Louise's house. There are three comparisons."

This is how Mrs. Jaeger herself explained her choice of the item:

> "The 'comparing, contrasting' was using a real life situation to compare the character in the story. The other situation 'comparing, contrasting' was put yourself in her shoes. Thinking, 'that's not how it is at my house.' That this is how Elizabeth's grandmother's house was. They had to of course, compare and contrast."

In light of this evidence, it is difficult to imagine any rater disagreeing that this segment of instruction was not in line with the glossary definition for this item:

> B1k, Comparing, contrasting - Include work on understanding text that requires the reader to tell how the story or characters or events or information is similar to or different from another part of this text or a different text. If the work requires the reader to compare something to a life event, record that under "making personal connections to text."

In explaining his response, the dissenting observer ma'e a fine distinction based on the definition stating that, "I did not mark this item because the target student was comparing the houses within a story not comparing aspects of two different stories." In this case, it appears that this observer made a more fine-grained distinction than warranted by the definition. The end result was that, even though this observer generally agreed with his fellow raters that "comparing, contrasting" occurred in this lesson segment, his response contributed to inter-rater disagreement.

At times, observers lacked crucial contextual information that teachers possessed.

In a number of cases, teachers' knowledge of their students, of the curriculum they were teaching, or of the instructional events that preceded the lesson that was observed, shaped their choice of log items. Observers' lack of access to this knowledge in these cases often led observers to choose different items than the teachers.

Two items in the "guided reading strategies" measure - B1d (Making personal connections to story or text) and B1e (Making predictions, generating questions) were sources of disagreement between Ms. Booth and her observers. Ms. Booth marked neither item, while both observers marked the former and one observer marked the latter. Ms. Booth explained her reasoning for not marking B1d in the interview:

> I go down that list but then I get makes personal connections with, to the story or textbook. This little reading group of mine, they make personal connections with everything. I have to always look that up and say that's not what they mean. When we start talking off task, go off on tangents, because that group does. I just mark the two. Locating answers and summarizing because that was the areas that I really worked on with him...I don't know. This part, if it was any of my other reading groups but the one. That was my gifted group. Those children are all gifted or talented [inaudible] and that's a whole different world. You heard the other reading groups, it's like a different world...
>
> I didn't mark it because according to what I saw in the glossary what they're emphasizing is they use the connection so they understand the story. That's not what they were doing. They were, do you understand what I'm saying? I try to be honest about it. I could've marked that oh, yeah, yeah, that's what they did but that's not what they did. I know what they did. I've been struggling with this little section for quite a few days...To further understand the story. They are not doing that. They understand the story. They are making personal connections because they just think at a different level. I hate to say it but that's what they're doing.

Ms. Booth makes it clear that she did not choose item B1d because having students make personal connections was not an instructional goal she was pursuing that day. Rather, when students made such connections they were veering off task. She explained that making this distinction was a recurring problem and partially attributed this problem to the ability level of these students. In formulating her response, Ms. Booth drew upon a rich store of contextual

information about the students she teaches, recurring interactions she has with those students, and her long and short term instructional goals. Very little if any of this information was available to the observers, and this asymmetry in contextual knowledge appears to have led raters to check different items. Furthermore, Ms. Booth also seems to imply that her responses to certain log items may be conditioned by the particular students on whom she is reporting. This possibility is troubling for a survey instrument that is designed to capture instructional practice in a standardized fashion because it suggests that Ms. Booth may vary her interpretation of and response to certain log items depending on which students she is reporting on.

In both these examples, teachers' possession of critical contextual information led them to believe that an instructional event was significant enough to warrant a check on the log. The way in which raters' perceptions about the significance of instructional events affected their responses is the focus of the next section.

At times, raters' interpretation of the significance of an instructional event differed. In follow-up interviews and in observers' reflective notes, several raters said they were uncertain whether there was "enough" of a particular instructional activity within a lesson to warrant a check mark on the log. In many of these cases, this uncertainty was reflected in inter-rater disagreement. There were a number of cases where very brief lesson segments were reported on the log. For example, both researchers who observed Ms. Temple checked B1k "comparing, contrasting" for the following exchange which lasted no more than a minute:

> Ms. Temple: Do you know what that just reminded me of?
>
> Student: The one with the meatballs.
>
> Ms. Temple: Remember when it got crazy in the town of ChewandSwallow? Food everywhere. Well, this is the same problem, too much food.

Unlike the observers, Ms. Temple did not mark B1k, probably because she felt that this was not a significant focus of the reading comprehension lesson, or perhaps because she simply forgot she had asked the question.

In a second case, raters' differing characterizations of a writing exercise within a comprehension lesson led them to code the lesson differently. In general, raters had to decide whether writing within a comprehension lesson was significant enough to qualify as "extended" (B2e) or whether it should be recorded as "brief" (B2c). The distinctions made in the glossary definitions appear quite clear on the surface:

> B2c. Provide brief written comments or answers - The target student answered questions requiring written responses less than a paragraph (e.g., a worksheet with questions about specific details, or questions placed on the board that were answered by writing a sentence or two).
>
> B2e. Provide extended written answers - The target student answered questions requiring extended responses (a paragraph or more in length). For example, a worksheet or response journal with questions requiring explanation, elaboration, or other involved responses.

Nevertheless, a number of raters had difficulty making this distinction. In Ms. Stenkl's class for example, the target student wrote the following sentences:

> Kevin is made.
> Kevin dus not like baby-sitters.
> Kevin does not whant his mom to leav.
> Kevin does not like kissy kissy books.
> Kevin likes baseball.

One observer viewed these as single sentences and subsequently coded this segment as involving "brief written comments or answers" (B2c). The observer wrote:

> "I called this brief since they were less than a sentence…I don't think that the teacher asked the target student for extended responses as defined in the glossary. However, the boundary between brief and extended is a fuzzy one, especially when students are not writing a great deal – that is young writers such as first grade students."

Meanwhile the other observer coded these sentences as "extended written answers" (B2e) and justified his choices as follows:

> "because the target student wrote several sentences which consisted, really, almost of a paragraph."

As the first observer indicates, assessing the complexity of a written exercise is not an entirely straightforward process, especially when early writers are involved. There is clearly a thread running through the sentences the target student wrote and one might construe those sentences as a paragraph about a boy named Kevin. This case suggests that even more specific coding definitions might be needed for some items. However, the fact that some raters seemed to pay very little attention to glossary definitions and that others who did use the glossary still interpreted items in different ways, casts doubt on whether this kind of change would translate into more accurate reporting.

One of the more difficult things we have struggled with on the instructional log is to help teachers distinguish between significant instructional events and events that are more superfluous. As the above examples illustrate, assessing the significance of an instructional task is not a simple problem. These examples and others like them suggest to us that some teachers may regularly report on very brief instructional segments while others may regularly filter such segments from their reporting, thinking that they are insignificant or superfluous. While these sorts of between-teacher differences may not have a tremendous impact on a report for a single day of instruction, cumulated over time as log reports are, these sorts of differences could significantly impact the validity of log measures.

At times, item redundancy facilitated inconsistent responses. Because the log measures very discrete aspects of instruction, there is built in redundancy in the instrument. For example, the same segment of instruction may be characterized by one rater as "Summarizing"

(B1g) while characterized by another rater as "Identifying main ideas or details" (B1n). Although the log glossary contains explicit instructions about when to choose these two items, the distinction between these two facets of instruction can be very fine. It is often difficult to decide which of these items to choose given an instructional segment. Take for example the glossary definition for "identifying main ideas and details" (B1n):

> Include work on understanding text that requires identifying main ideas and details such as setting, names of characters, or details included in a non-fiction selection. This category could also include a student's retelling of a story or text. However, if the purpose of the retelling is to identify only the critical details, please record the activity under summarizing.

If in retelling a story a student identified one critical detail and a number of inconsequential details, would selecting this item be appropriate? Is there a certain proportion of items that would need to be "critical" before a rater could code a lesson segment as "summarizing" instead of "identifying main ideas and details"? Our analyses suggested that having items on the log that separate instruction into such subtle shades often led directly to disagreement among raters.

In a considerable number of cases, raters marked different, but conceptually similar items for the same segment of instruction. When the items chosen by the raters are in the same measure, the measurement disagreement index is unaffected but the item disagreement index is increased. This is what happened in the observation for Ms. Becker. Both Ms. Becker and the observer marked three of the same items within the "guided comprehension" measure. However, Ms. Becker also marked B1a "Activating prior knowledge", while the observer marked B1d "Making personal connections to story or text." These items are similar and this is acknowledged in the observer's rationale for why he mismatched with Ms. Becker:

> Ms. Becker marked this item (B1a) and I didn't. I believe she marked this because of the newspaper article she handed out to the Ss to read before they read Tornado Alert. They talked a little bit about things they had already learned about Tornadoes, such as Tornado

> alley and what states may be part of Tornado alley. I think this is a cross with B1d. I marked B1d instead of B1a because I considered the newspaper article to be text in itself and the questions the teacher asked followed the Ss reading of the article.

There is further evidence in the follow-up interview with Ms. Becker that the reason she marked B1a "Activating prior knowledge" instead of B1d "Making personal connections" was because they had discussed the material she was asking questions about on a previous day, which was critical contextual information that the observer was not aware of.

> Yeah, remembering what I said. I'm trying to think whether or not, most days I do whether or not we talked a little bit about what we had read about or the storms that had taken place previously the day before. I think we talked about as we went or before hand. To activate their knowledge...You know what, I did my log strictly on the "Tornado Alert." I did. Later, I didn't. But, as I started it that's where I was. I know we had talked about the storms in Nebraska the day before.

It appears, therefore, that the teacher and the observer really had the same incident in mind when they marked different, yet conceptually similar items. Glossary instructions for item B1a "Activating prior knowledge" state, "If this activity occurs after having read the text, record it as, 'making personal connections to story or text'." Therefore, the items are intended to mean similar things and the confusion between the observer and the teacher can apparently be attributed to the fact that the observer did not know that the subject of tornadoes had been discussed the previous day. Because the items tap such similar things, and because the two items are in the same measure, the impact of on the meaning of Ms. Becker's score on the "guided comprehension" measure seems fairly inconsequential.

What happens when raters choose two different but similar items that are not in the same measure? For example, the items B1c "Locating answers or information" which is in the "Guided comprehension" measure and B1n "Identifying main ideas or details" which is in the "Summarizing/describing what was read" were often used interchangeably by raters. When raters choose different but conceptually similar items that are in two different measures, scores

on the measure disagreement index and item disagreement index for both measures will be increased.

Raters also mismatched on a number of items that were logically linked to one another, and therefore, redundant to a certain degree. For example, the items B1g "Summarizing", B2o "Retell story", and B2p "Summarize text" are logically linked to one another. Specifically, in working on the comprehension area of summarizing (B1g), students are frequently asked to either retell a story (B2o) or summarize text (B2p). Raters often recognized the logical connection between these items in their responses. However, when one rather checked the logically-connected pairs of items and other raters did not, mismatches occurred.

There was also a logical connection between the set of comprehension strategies contained in items B1a through B1i, and an item that indicated whether students were practicing such strategies (B2j, practice teacher-selected comprehension strategies). This occurred in Ms. Antos' observation. Observers' notes from that day indicate that the key comprehension activity was the strategy "previewing, surveying" (B1f). Ms. Antos and both observers checked this item. In addition, the observers checked B2j to indicate the students were practicing "previewing and surveying" but Ms. Antos did not check this item. Unfortunately, there is no evidence from Ms. Antos' follow-up interview about why she did not choose B2j.

In Ms. Antos' case, the impact of this mismatch on the meaning of her proxy measure scores seems modest. Again, the key activity that day appeared to be "previewing, surveying" and Ms. Antos checked that item. Item B2j largely functions as a confirmation that students were in fact practicing the comprehension strategy "previewing, surveying", but that fact might reasonably be inferred from Ms. Antos' selection of item B1f.

Summary and Discussion

This series of three different sets of analyses progressively moved us closer to understanding why the three measures were or were not valid. The first two analyses really provided us with gross diagnostic information. The factor analyses allowed us to evaluate the construct validity of the measures by demonstrating which items fit together and which items did not. These analyses also provide us a traditional measure of validity, the internal consistency of respondents' answers. The analyses of the inter-rater agreement indices alerted us to the fact that there was substantial disagreement among raters. But beyond this rather superficial piece of information, these results also provided very detailed evidence about the nature and the magnitude of inter-rater disagreement. For example we learned from these analyses that teachers generally reported much higher levels of reading comprehension activity than observers and that on average, raters marked different items about one quarter of the time. The qualitative evidence we examined identified a number of specific reasons why raters disagreed. Some of these, such as differences that are attributable to asymmetries in the contextual knowledge held by raters strike us as problems that are inherent in measuring instruction that are not going to be susceptible to quick fixes. Others, however, like item redundancy, are things that are clearly under the control of researchers and are thus things that can be changed to increase the validity of measures. In fact, a number of changes that have grown out of the V*alidity Study* have already been incorporated into a modified log.

Our primary vehicle for untangling the validity of the measures was to assess inter-rater agreement. Along the way we gained a deeper understanding of what it actually means to be high or low on our measures. In our minds this is the ultimate goal of a validity analysis and is something we intend to pursue further with this instrument. Examining the qualitative evidence,

we saw rich snapshots of instruction that was often reflected quite well by teachers' log reports. We also saw a number of cases where we wondered whether the answers a teacher provided on the log captured what she was actually doing in her classroom with her students. Taken as a whole, the three sets of analyses provided us with rich diagnostic information that has allowed us to improve the log and its accompanying procedures and materials. We could not have learned all we did had we used the three approaches in isolation from one another. In fact, had we used the three approaches in isolation, we might have pursued three different, possibly contradictory, courses of action. All our problems have not been solved, in fact, these analyses certainly unearthed a number of problems we never expected to find. But this work has reinforced our belief that knowing is better than not knowing, and knowing more is better than knowing less.

References

Barnes, C., Correnti, R., and Taylor, J. (2001). Marshalling evidence for validity: Problems and implications for measuring instruction. Paper presented at the 2001 Annual Meeting of the American Educational Research Association in Seattle, WA.

Fee, J. (1979). Symbols and attitudes. Unpublished doctoral dissertation. University of Chicago.

Freeman, D. (1996). To take them at their word: Language data in the study of teachers' knowledge. Harvard Educational Review. 66(4).

Jick, T. (1979) Mixing qualitative and quantitative methods: Triangulation in action. Administrative Science Quarterly, 24, 602-611.

Kim, J-O., and Mueller, C.W. (1978). Factor analysis: Statistical methods, and practical issues. Newberry Park, CA: SAGE.

Lackner, J.R., and Garrett, M.F. (1972). Resolving ambiguity: Effects of biasing context in the unattended ear. Cognition. 1.

LeCompte, M. and J. Goetz (1982) Problems of reliability and validity in ethnographic research. Review of Educational Research, 52 (1), 31-60.

Moss, P. A. (1994) Can there be validity without reliability? Educational Researcher, 23 (2), 5-12.

Sudman, S., Bradburn, N.M., and Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass.

Appendix A - Language Arts Log

LANGUAGE ARTS LOG

Carefully place your student label here

1. **Did you teach reading or language arts today?** Mark (X) ONE box.
   - ☐ Yes
   - ☐ No - *Skip to the Comments box at the end of Section C*

2. **How much time did the student spend in reading or language arts today?** Print number of minutes in the boxes

   ☐☐☐ **Minutes** If less than 100 minutes, please place zero ("0") in first box.

3. **How much of this time did the student actively work on assigned tasks?** Mark (X) ONE box.
   - ☐ most of the time
   - ☐ some of the time
   - ☐ a small amount of the time
   - ☐ none of the time

4. **What percentage of this Language Arts time was instruction provided by...** Mark (X) ONE box.

   Print percent in the boxes below.

   ☐☐☐ **You** If less than 100 percent, please place zero ("0") in first box.

   ☐☐☐ **Another teacher** If less than 100 percent, please place zero ("0") in first box.

   ☐☐☐ **Aide** If less than 100 percent, please place zero ("0") in first box.

5. **To what extent were the following topics a focus of the student's work in reading/language arts today?** Mark (X) EACH item below, 5a-5h.

	Primary Focus	Secondary Focus	Touched on only Briefly	Not a Focus	Complete section(s) if this topic was the PRIMARY or SECONDARY focus
a. Word analysis.....	☐	☐	☐	☐	A
b. Comprehension.....	☐	☐	☐	☐	B
c. Writing.....	☐	☐	☐	☐	C
d. Concepts of print.....	☐	☐	☐	☐	None
e. Vocabulary.....	☐	☐	☐	☐	None
f. Research strategies.....	☐	☐	☐	☐	None
g. Grammar.....	☐	☐	☐	☐	None
h. Spelling.....	☐	☐	☐	☐	None

If you marked "Primary Focus" or "Secondary Focus" for Questions 5a, 5b, or 5c, turn the page and answer the questions for the section(s) indicated in the colored boxes above.
All others go to the Comments box at the end of Section C.

SECTION A - Word Analysis

A1. What areas of word analysis did you work on with the student today?
Mark (X) EACH area that you worked on

- ☐ Letter-sound relationships (A1a)
- ☐ Sound segmenting (A1b)
- ☐ Sound blending (A1c)
- ☐ Sound writing, sound spelling (A1d)
- ☐ Word families, phonograms (A1e)
- ☐ Word recognition, sight words (A1f)
- ☐ Structural analysis (A1g)
- ☐ Use of context, picture, syntactical cues to read words (A1h)
- ☐ Use of phonics-based cues to read words (A1i)

A2. What did you have the student do in word analysis today?
Mark (X) EACH item that you had the student work on.

Listen for a sound, and . . .

- ☐ circle or point to the letter that makes that sound (A2a)
- ☐ say the name of the letter that represents that sound (A2b)
- ☐ write the letter that represents that sound (A2c)

- ☐ Give the sound for a written letter (A2d)
- ☐ Listen for sounds in words (A2e)
- ☐ Look for words in text that have a particular sound or sound pattern (A2f)
- ☐ Produce words that contain a given sound (A2g)
- ☐ Recognize or use structural relationships - word families, common syllables (A2h)
- ☐ Blend sounds from oral teacher prompt (A2i)
- ☐ Blend sounds using written text (A2j)
- ☐ Segment words into sounds - orally (A2k)
- ☐ Write the sounds heard in words (A2l)
- ☐ Memorize, identify or read sight words (A2m)
- ☐ Complete sentences with correct word - cloze (A2n)
- ☐ Use picture cues to read (A2o)
- ☐ Use context cues to read (A2p)
- ☐ Use phonics cues to read (A2q)
- ☐ Choral or echo read (A2r)
- ☐ Reread for practice and fluency (A2s)

A3. What materials did the target student use?
Mark (X) EACH item used by the student.

- ☐ Pictures or objects to identify sounds, letters, words (A3a)
- ☐ Isolated words or letters (A3b)

Individual sentences . . .

- ☐ with picture cues (A3c)
- ☐ without picture cues (A3d)

Connected text . . .

- ☐ with controlled sight vocabulary (A3e)
- ☐ with patterned or predictable language (A3f)
- ☐ with predominant phonetic pattern (A3g)
- ☐ that is literature-based or thematic (A3h)

A4. How did you interact with the target student?
Mark (X) EACH item that applies.

- ☐ I explained or modeled how to read, write, or identify letters, sounds or words (A4a)
- ☐ I mixed an explanation with brief questions to students (A4b)

I helped the student practice . . .

- ☐ correcting his or her errors (A4c)
- ☐ answering the student's questions (A4d)
- ☐ asking questions (A4e)
- ☐ giving oral cues (A4f)
- ☐ prompting for strategy use (A4g)

- ☐ I read to or with the student (A4h)
- ☐ I listened to the student read (A4i)
- ☐ I took running records or conducted a miscue analysis (A4j)
- ☐ I used a scripted lesson (A4k)
- ☐ I administered a test (A4l)
- ☐ The student worked independently without my assistance (A4m)

Proceed to Section B and/or C, <u>ONLY IF</u> you marked "Primary Focus" or "Secondary Focus" at Questions 5b or 5c. All others turn the page and go to the Comments box at the end of Section C.

2247

Office Use ONLY

SECTION B - Comprehension

B1. What areas of comprehension did you work on with the student today?
Mark (X) EACH area that you worked on.

Approaches to understanding text

- ☐ Activating prior knowledge (B1a)
- ☐ Clarifying, monitoring for meaning (B1b)
- ☐ Locating answers or information (B1c)
- ☐ Making personal connections to story or text (B1d)
- ☐ Making predictions, generating questions (B1e)
- ☐ Previewing, surveying (B1f)
- ☐ Summarizing (B1g)
- ☐ Using concept maps, text structure frames (B1h)
- ☐ Using specific strategies - e.g., KWL, SQ3R (B1i)

Ways of demonstrating understanding of text

- ☐ Author's craft (B1j)
- ☐ Comparing, contrasting (B1k)
- ☐ Critical stance, analyzing and evaluating text (B1l)
- ☐ Drawing conclusions, making inferences (B1m)
- ☐ Identifying main ideas and details (B1n)
- ☐ Interpreting text aides - e.g., charts, graphs (B1o)
- ☐ Story structure (B1p)

Extension/practice

- ☐ Literature extension activities (B1q)
- ☐ Reading for pleasure or information (B1r)

B2. What did you have the target student do in comprehension today?
Mark (X) EACH item that you had the student work on.

- ☐ Listen to text read to them (B2a)

Provide brief comments or answers

- ☐ Oral (B2b)
- ☐ Written (B2c)

Provide extended answers

- ☐ Oral (B2d)
- ☐ Written (B2e)

- ☐ Work on a written project (B2f)
- ☐ Work on a creative (non-written) literature extension activity (B2g)
- ☐ Practice specific skills - e.g., draw conclusions, identify similes (B2h)

Practice comprehension strategies

- ☐ Student selected (B2i)
- ☐ Teacher selected (B2j)

- ☐ Explain application of skills or strategies (B2k)
- ☐ Conduct a think aloud (B2l)
- ☐ Frame and ask questions about text (B2m)

B2. What did you have the target student do in comprehension today? (continued)

- ☐ Justify answers or explain reasoning (B2n)
- ☐ Retell story (B2o)
- ☐ Summarize text (B2p)
- ☐ Create a story frame or text structure map (B2q)
- ☐ Complete sentences with correct word or words (B2r)

Read. . .

- ☐ Assigned text (B2s)
- ☐ Student chosen text (B2t)

B3. What materials did the target student use?
Mark (X) EACH item used by student.

- ☐ Informational text (B3a)

Narrative text . . .

- ☐ with controlled sight vocabulary (B3b)
- ☐ with patterned or predictable language (B3c)
- ☐ with a predominant phonetic pattern (B3d)

Literature based or thematic . . .

- ☐ short selection (B3e)
- ☐ chapter book (B3f)

B4. How did you interact with the target student?
Mark (X) EACH item that applies.

- ☐ I explained a reading strategy or conducted a think aloud (B4a)
- ☐ I led a discussion (B4b)

I read to the target student individually or in a group...

- ☐ with only minimal comments or questions (B4c)
- ☐ frequently asking questions about text / commenting on text (B4d)
- ☐ encouraging students to pose questions / suggest strategies (B4e)

I listened to the target student read . . .

- ☐ asking minimal comments or questions (B4f)
- ☐ frequently asking questions about text / commenting on text (B4g)
- ☐ encouraging the student to pose questions / suggest strategies (B4h)

I monitored the student's work or other extension activities . . .

- ☐ answering the student's questions (B4i)
- ☐ asking student to explain reasoning (B4j)
- ☐ asking questions to prompt student thinking (B4k)
- ☐ correcting the student's errors (B4l)
- ☐ commenting on the student's work (B4m)
- ☐ prompting the student for strategy use (B4n)

- ☐ I used a scripted lesson (B4o)
- ☐ I administered a test (B4p)
- ☐ The student worked independently without my assistance (B4q)

SECTION C - Writing

C1. What areas of writing did you work on with the student today?
Mark (X) for EACH area that you worked on.

- ☐ Generating ideas for writing (C1a)
- ☐ Organizing ideas for writing (C1b)
- ☐ Literary techniques, author's style (C1c)
- ☐ Writing forms or genres - e.g., letter, drama, editorial (C1d)
- ☐ Writing practice (C1e)
- ☐ Revision of writing (C1f)
- ☐ Editing (C1g)

C2. What did you have the target student do in writing today?
Mark (X) EACH item that you had the student work on.

- ☐ Generate ideas for writing (C2a)
- ☐ Organize ideas for writing (C2b)
- ☐ Free write (C2c)
- ☐ Write using literary techniques (C2d)
- ☐ Imitate writing style of a published author (C2e)
- ☐ Write in a structured format, such as business letter, poetry (C2f)
- ☐ Write for a particular audience (C2g)
- ☐ Other teacher assigned writing (C2h)
- ☐ Revise writing (C2i)
- ☐ Edit writing (C2j)
- ☐ Examine writing style of a published author (C2k)
- ☐ Observe teacher lesson (C2l)
- ☐ Share their writing with others (C2m)

C3. The target student wrote . . .
Mark (X) EACH item that applies

- ☐ letter strings or words - with or without illustration (C3a)
- ☐ separate sentence(s) - with or without illustration (C3b)
- ☐ separate paragraph(s) (C3c)
- ☐ connected paragraphs (C3d)

C4. How did you interact with the target student?
Mark (X) EACH item that applies.

- ☐ I demonstrated or did a think-aloud using my own writing (C4a)

I explained how to write, organize ideas, revise or edit ...

- ☐ using student writing (C4b)
- ☐ using a published author's writing (C4c)
- ☐ I took dictation from the student (C4d)
- ☐ I led the student and his/her peers in a group composition (C4e)
- ☐ I acted as an audience for student writing -commenting on content of the writing (C4f)

I monitored seatwork . . .

- ☐ commenting on positive aspects of student work (C4g)
- ☐ asking questions (C4h)
- ☐ prompting student to elaborate (C4i)
- ☐ prompting student to refine or reorganize (C4j)
- ☐ helping student with spelling and punctuation (C4k)
- ☐ helping student with grammar and syntax (C4l)

The student worked independently

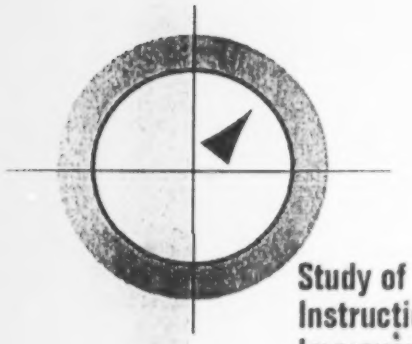
- ☐ with a writing or proofreading guide (C4m)
- ☐ without any written guide (C4n)

COMMENTS - Please PRINT clearly..Thanks

2247

Office Use ONLY

Appendix B - Language Arts Log Glossary



Study of Instructional Improvement

LANGUAGE ARTS GLOSSARY

Institute for Social Research
University of Michigan
Ann Arbor

Toll-free number for SII-ISR Office:
1-877-397-2374
February 2000

Language Arts Glossary
Section B: Comprehension

B1. What areas of comprehension did you work on with the student today?
Please check all the areas within comprehension that you and the target student worked on today. Include both listening and reading comprehension.

Approaches to understanding text

B1a. Activating prior knowledge
Include activities in which you demonstrated or asked the target student to consider what they already know about a topic as it relates to the content of the text prior to reading the text. If this activity occurs after having read the text, record it as "Making personal connections to story or text".

B1b. Clarifying, monitoring for meaning
Include activities where you demonstrated or asked the target student to check understanding as they read. For example, does what they are reading make sense to them? Or, you may have had the target student check understanding of difficult vocabulary and unfamiliar concepts; or you may have taught students to identify dense or complicated text and modeled how to slow down reading speed to allow for better understanding.

B1c. Locating answers or information
Include activities in which you demonstrated how to find answers or you guided the target student in how to locate answers to different kinds of questions. You may have also required the target student to explain how to locate answers to different kinds of question. Answers might require students to locate details in text or synthesize information; or answers might not be directly in the text, requiring students to call upon background knowledge.

B1d. Making personal connections to story or text
Include activities in which you ask the target student to consider the text in relation to personal experiences, opinions, or background knowledge.

B1e. Making predictions, generating questions
Include activities in which you demonstrated or asked the target student to set a purpose for reading, making predictions about what would be in the text, or formulating questions that you expect to have answered by the text. Questions would be generated before reading that section of text. You might have asked the target student to generate questions from what they learned in previewing text and/or from prior knowledge.

B1f. Previewing, surveying
Include activities in which you demonstrated or asked the target student to quickly survey a text to determine what they think the text is about, and/or explain how previewing or surveying before reading helps in reading and understanding text.

B1g. Summarizing
Include activities in which you demonstrated or asked the target student to state the important information learned in text summarizing key concepts, or to recite answers to questions generated in pre-reading. If the student is doing a simple retelling of the information he/she remembers from the story and is not working on summarizing the critical aspects, then record the activity as working on "identifying main idea and details."

B1h. Using concept maps, text structure frames
Include activities in which you demonstrated the use of visual organizers to comprehend text, or in which you provided visual organizers for the target student to use. For example, you may have modeled how to use "pro and con charts," where you chose a claim or statement and listed facts and details that supported or contradicted this claim. Other examples include compare and contrast charts, cause and effect diagrams, timelines, webs, outlines, or story frames.

B1i. Using specific strategies --e.g., KWL, SQ3R
Include demonstrating or practicing specific strategies (i.e., established methods for making sense of stories or text that include a standard and shared procedure or set of steps that can be taught and learned). Some examples include KWL, Reciprocal Teaching, SQ3R, SAIL, QAR. Check this category only if you taught (demonstrated or explained) or practiced all the steps in a given strategy. Mark in previous categories any parts of a strategy that you demonstrated or practiced.

Ways of demonstrating understanding of text

B1j. Author's craft
Include work on understanding text that requires the reader to stand apart from the text and consider the techniques the author used in telling the story. For example: "How did the author create a suspenseful mood?"; "Why did the author begin with a description of the house?"; "How did the author let you know how the main character was feeling?"

B1k. Comparing, contrasting
Include work on understanding text that requires the reader to tell how the story or characters or events or information is similar to or different from another part of this text or a different text. If the work requires the reader to compare something to a life event, record that under "making personal connections to text."

B1l. Critical stance, analyzing and evaluating text
Include work on understanding text that requires the reader to stand apart from text and to consider it critically. The text may be fiction or nonfiction. This category

includes questions about the adequacy of evidence and consistency of reasoning. For example: "How believable is the ending?"; "What were the clues that led to this conclusion?"; "Was there enough information for us to believe that the main character could act in this way?", "Is the author's argument supported by facts?".

B1m. Drawing conclusions, making inferences
Include work on understanding text that requires drawing conclusions or making inferences based on the evidence and information in the story. For example, this category includes using information from the story to explain how a character will act or react, identifying character traits that are implied in the story, or drawing conclusions based on facts presented in a nonfiction selection.

B1n. Identifying main ideas and details
Include work on understanding text that requires identifying main ideas, and details such as setting, names of characters, or details included in a nonfiction selection. This category could also include a student's retelling of a story or text. However, if the purpose of the retelling is to identify only the critical details, please record the activity under "summarizing".

B1o. Interpreting text aides - e.g., charts, graphs
Include work on using text aids such as maps, tables, charts, figures to locate or interpret information or details.

B1p. Story structure
Include work on understanding text that requires understanding the structure of story (e.g., identifying the problem, sequencing the plot, and explaining story resolution).

Extension/practice

B1q. Literature extension activities
Include work on an activity that extended understanding of a text selection or story. For example, the target student may have created puppets to re-enact a story.

B1r. Reading for pleasure or information
Include interactions in which students are asked to read a story or passage for a sustained period of time. The target student may read silently or aloud. The text may be chosen by the target student or by the teacher. Examples of reading times include partner reading, SSR (sustained silent reading) and DEAR (Drop Everything and Read).

B2. What did you have the target student do in comprehension today?

B2a. Listen to text read to them
The target student listened to a narrative story or passage read aloud by you (the teacher), by another student, or on an audio tape.

Provide brief comments or answers

B2b. Oral
The target student answered questions orally. The questions required answers that could be given in a few sentences (e.g., the target student identified the main idea or described the major trait of a main character).

B2c. Written
The target student answered questions requiring written responses less than a paragraph (e.g., a worksheet with questions about specific details, or questions placed on the board that were answered by writing a sentence or two).

Provide extended answers

B2d. Oral
The target student answered questions orally that required extensive answers (e.g., more than a few sentences). For example, the target student might explain parallels between this story and a story previously read or explain evidence that led to assigning certain traits to a character.

B2e. Written
The target student answered questions requiring extended responses (a paragraph or more in length). For example, a worksheet or response journal with questions requiring explanation, elaboration, or other involved responses.

B2f. Work on a written project
The target student extended their understanding of text through a written project. For example, the target student may have rewritten the story from a different point of view, written a new ending, written a letter to the author, written a review of the book for the school newspaper, or rewritten a story into a play.

B2g. Work on a creative (non-written) literature extension activity
The target student worked on creating a response to a text selection using a medium other than writing. For example, the target student made a shadow box of an event in the story, acted out a play based on the text, created puppets and had the puppets act out the story, drew an illustration of a concept or event, or created a game based on information in the text.

B2h. Practice specific skills - e.g., draw conclusions, identify similes
The target student answered questions that required the target student to utilize skills such as drawing conclusions, sequencing events, identifying main characters or identifying the main idea.

Practice comprehension strategies.

B2i. Student selected
The target student selected the strategies they would use in understanding text. For example, one student might have chosen to generate questions before reading, while another student might have chosen to use a graphic aid to help in understanding text.

B2j. Teacher selected
The teacher assigned a strategy for the target student to practice, e.g., surveying the text, making predictions, generating questions.

B2k. Explain application of skills or strategies
The target student explained the steps in applying a skill or strategy to text. This could occur prior to using the skill or strategy, or after applying the skill or strategy. The target student might have done this orally or in writing.

B2l. Conduct a think aloud
The target student orally discussed or explained their thinking about and attempts to understand a story or text as they read it.

B2m. Frame and ask questions about text
The target student asked questions about the meaning of the content of the text (e.g., "Why did [main character] do that?"). The questions may have been framed, pursued, and answered only by the target student, or they may have also been directed at the teacher or to other students in the class. The questions may have been formulated in pre-reading or may have been formulated after reading.

B2n. Justify answers or explain reasoning
The target student used text to justify answers given to comprehension questions or explained their reasoning.

B2o. Retell story
The target student generated an account of a story that includes both important (major) and minor details. If the target student provides a succinct account of the main features of a story, record this activity in "Summarize text."

B2p. Summarize text
The target student reviewed information learned in text or reviewed answers to questions posed before or during reading, or summarized a story by providing a succinct account of the main features of a story.

B2q. Create a story frame or text structure map
The target student completed a visual organizer such as a story frame, web, character map, concept map, or outline of the story. The organizer may have been created by the target student, or provided by the teacher and completed by the target student. These organizers usually are designed to help the target student understand the sequence of events, major parts of the story, or relationships in a story.

B2r. Complete sentences with correct word or words
The target student read a sentence (or sentences) missing one or more key words or concepts that had been presented in a larger text. The target student completed the sentence(s) with the correct word or words. This activity is sometimes referred to as a cloze procedure.

Read . . .

B2s. Assigned text
The target student read stories, articles, poems or other text that was assigned by the teacher or as a part of the curriculum.

B2t. Student chosen text
The target student read stories, articles, poems or other text that was self-selected by the student. The target student may have made a choice from among a selection of two or more teacher recommended texts.

B3. What materials did the target student use?

B3a. Informational text
Informational or expository text provides information, instructions, or facts. For example, many science articles are informational. Directions and recipes are also informational. Include in the informational category, text designed to inform, instruct or persuade.

Narrative text . . .
Narrative text refers to stories. For example, novels, fables, fairy tales, and children's stories are all considered narratives. Many poems and plays are also narratives. Typically, narratives have characters and plots. Include text that has a primary purpose of entertainment or enjoyment such as poems, songs, and rhymes in the narrative text category.

B3b. with controlled sight vocabulary
The target student used text that is intentionally designed to be easier or more accessible for beginning readers. Vocabulary may be controlled by: 1) including only words that are phonetically regular using patterns previously taught; 2) using a high number of common sight words (e.g., this, the, boy, look), or; 3) adding very few new irregular words (i.e., words not easily decodable) and introducing

them in context. If a strong phonetic pattern is evident in the text (e.g., many words with a short /a/vowel sound "the cat sat on the bag"), use the category "with a predominant phonetic pattern ".

B3c. with patterned or predictable language
The text used by the target student included repeated patterns and/or refrains, i.e., rhyming patterns or verses repeated throughout (e.g., poems or text such as Brown Bear, Brown Bear, Napping House, or If You Give a Mouse a Cookie).

B3d. with predominant phonetic pattern (B3d)
The target student used text with a majority of the words following a strong phonetic pattern (e.g., all short a words, or many words having a vowel-consonant-vowel pattern). The remainder of the words in the text are usually within the sight vocabulary of the target student or are introduced prior to working with the text.

Literature based or thematic . . .
The target student read stories or trade books containing subject matter that is understandable to children of this age, however, no organized effort was made to strategically limit the words used in the text.

B3e. short selection
This category includes illustrated story books, short stories, articles, and poems. A single selection within an anthology or a basal reader would also be included in this category.

B3f. chapter book
This category includes books that have multiple connected chapters. Mark this category even if the target student only read one chapter, or a part of a single chapter, on a given day as long as the remainder of the book is read at other times.

B4. How did you interact with the target student?

B4a. I explained a reading skill or strategy or conducted a think aloud
Include interactions in which you demonstrated a skill or strategy, or explained the steps in a skill or strategy, or told the target students how you decided to apply a skill or strategy as you read the text (i.e., conducted a think-aloud).

B4b. I led a discussion
Include interactions in which you made comments about the story or text, asked questions about the story or text, and/or asked the target student questions about how to apply a strategy or skill.

I read to the target student individually or in a group...
Include the interactions in which you personally read aloud to the target student either individually or in a group setting. Do not include taped readings.

B4c. with only minimal comments or questions
Include interactions in which you read without asking any questions or commenting, or when you interrupt the reading with only 1 to 3 comments or questions.

B4d. frequently asking questions about text or commenting on text
Include interactions in which you frequently stop your reading to ask questions or comment on text.

B4e. encouraging students to pose questions and suggest strategies
Include interactions in which you involve the target students in posing questions or suggesting strategies about the text you are reading.

I listened to the target student read . . .
Include interactions in which you listened to the target student read aloud either to you individually or in a group setting.

B4f. asking minimal comments or questions
Include interactions in which you listened to the target student read without commenting or asking any questions, or when you interrupt the reading with only 1to 3 comments or questions.

B4g. frequently asking questions about text / commenting on text
Include interactions in which you frequently stop the target student to ask questions or comment on text as you listen to him/her read.

B4h. encouraging the student to pose questions / suggest strategies
Include interactions in which you listen to the target student read and involve other students in posing questions or suggesting strategies about the text the target student is reading.

I monitored the student's work or other extension activities . . .
Include activities where you supervised the target student as he or she worked on answering questions, applying skills or strategies, or on literature extension activities (e.g., re-writing the story, creating a puppet show).

B4i. answering the student's questions
Include interactions in which you directly answered the target student's questions.

B4j. asking student to explain reasoning
Include interactions in which you asked the target student to provide justification or explanation for answers, opinions, or thoughts about text. For example, you may have asked how the student knows something, or what led to answering in a certain way.

B4k. asking questions to prompt student thinking
Include interactions in which you asked the target students questions to guide their thinking about how to respond to questions, how to understand a story or text, and/or to recognize errors in their responses to questions.

B4l. correcting the student's errors
Include interactions in which you corrected errors the target student made as she or he read text, worked on responding to questions, or practiced a skill or strategy.

B4m. commenting on the student's work
Include interactions in which you praised or commented on the target student's work without correcting or asking questions.

B4n. prompting the student for strategy use
Include interactions in which you suggested a student try a particular strategy or when you asked a question about which strategy might be helpful to use in understanding this text.

B4o. I used a scripted lesson
Include interactions in which you closely followed scripted instructions for teaching comprehension to the target student. Teaching scripts typically include a teaching prompt, expected responses, and reinforcement. Scripts may also include signaling or instructions for correction.

B4p. I administered a test
Include interactions in which you gave the target student a test, either written or oral. You may have tested the target student in any size group (individual, small group, or large group). The questions may have been multiple choice, short answer, or open-ended. You would not provide feedback to the target student until after the test was completed.

B4q. The student worked independently without my assistance
Include interactions in which the target student worked independently and you either worked with other students or worked on administrative tasks, but did not work directly with the target student. The target student may have received assistance from peers or from an aide, but during this time had no interaction with you other than behavioral guidance (keeping the student on task).